scrubadub Documentation

Release 0.0.1

Dean Malmgren

April 06, 2015

Contents

> **Warning:** This package is a work in progress and is not yet available on pypi. This documentation should be considered more of a design document for what scrubadub will do someday rather than a specification of what it can do today.

Remove personally identifiable information from free text. Sometimes we have additional metadata about the people we wish to anonymize. Other times we don't. This package makes it easy to seamlessly scrub personal information from free text, without comprimising the privacy of the people we are trying to protect.

`scrubadub` currently supports removing:

- names
- email addresses

CHAPTER 1

Quick start

Getting started with scrubadub is as easy as pip install scrubadub and incorporating it into your python scripts like this:

```
>>> import scrubadub

# John may be a cat, but he doesn’t want other people to know it.
>>> text = "John is a cat"

# Replace names with {{NAME}} placeholder. This is the scrubadub default
# because it maximally omits any information about people.
>>> placeholder_text = scrubadub.clean_with_placeholders(text)
>>> placeholder_text
"{{NAME}} is a cat"
```

CHAPTER 2

Related work

scrubadub isn't the first package to attempt to remove personally identifiable information from free text. There are a handful of other projects out there that have very similar aims and which provide some inspiration for how scrubadub should work.

- MITRE gives the ability to replace names with a placeholder like [NAME] or alternatively replace names with fake names. last release in 8/2014. not on github. unclear what language although it looks like python. it is clear that the documentation sucks and is primarily intended for academic audiences (docs are in papers).
- physionet has a few deidentification packages that look pretty decent but are both written in perl and require advance knowledge of what you are trying to replace. Intended for HIPAA regulations. In particular, deid has some good lists of names that might be useful in spite of the fact it has 5k+ lines of gross perl.

Contents:

2.1 Contributing

The overarching goal of this project is to remove personally identifiable information from raw text as reliably as possible. In practice, this means that this project, by default, will preferentially be overly conservative in removing information that might be personally identifiable. As this project matures, I fully expect the project to become ever smarter about how it interprets and anonymizes raw text.

Regardless of which peraonl information is identified, this project is committed to being as agnostic about the manner in which the text is anonymized, so long as it is done with rigor and does not inadvertantly lead to improper anonymization. Replacing with placholders? Replacing with anonymous (but consistent) IDs? Replacing with random metadata? Other ideas? All should be supported to make this project as useful as possible to the people that need it.

Another important aspect of this project is that we want to have extremely good documentation and source code that is easy to read. If you notice a type-o, error, confusing statement etc, please fix it!

2.1.1 Quick start

1. Fork and clone the project:

   ```
   git clone https://github.com/YOUR-USERNAME/scrubadub.git
   ```

2. Create a python virtual environment and install the requirements

   ```
   mkvirtualenv scrubadub
   pip install -r requirements/python-dev
   ```

3. Contribute! There are several open issues that provide good places to dig in. Check out the contribution guidelines and send pull requests; your help is greatly appreciated!

4. Run the test suite that is defined in `.travis.yml` to make sure everything is working properly

   ```
   ./tests/run.py
   ```

   Current build status:

2.2 Change Log

This project uses semantic versioning to track version numbers, where backwards incompatible changes (highlighted in **bold**) bump the major version of the package.

2.2.1 latest changes in development for next release

2.2.2 0.1.0

- initial release, ported from past projects

CHAPTER 3

Indices and tables

- *genindex*
- *modindex*
- *search*